ASSESSING PSYCHOLOGICAL TRAUMA AND PTSD

Edited by
JOHN P. WILSON
TERENCE M. KEANE

Foreword by Susan D. Solomon

THE GUILFORD PRESS
New York London

A Division of Guilford Publications, Inc.
72 Spring Street, New York, NY 10012

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Assessing psychological trauma and PTSD / edited by John P. Wilson and Terence M. Keane.
p. cm.
Includes bibliographical references and index.
ISBN 1-57230-162-7
1. Post-traumatic stress disorder—Diagnosis. 2. Psychodiagnostics. 3. Neuropsychological tests. I. Wilson, John P. (John Preston) II. Keane, Terence Martin.
[DNLM: 1. Stress Disorders, Post-Traumatic. WM 170 A846 1997]
RC552.P67A85 1997
616.85'21—dc20
DNLM/DLC
for Library of Congress 96-30206
CIP

CHAPTER 4

Psychometric Theory in the Development of Posttraumatic Stress Disorder Assessment Tools

FRANK W. WEATHERS
TERENCE M. KEANE
LYNDA A. KING
DANIEL W. KING

INTRODUCTION

In the last 15 years, scientific research on posttraumatic stress disorder (PTSD) has increased dramatically, yielding a wealth of knowledge regarding the clinical phenomenology, etiology, and treatment of this complicated, intractable disorder. Clearly the recognition of PTSD as a formal diagnostic entity in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III; American Psychiatric Association, 1980) provided considerable impetus for this extraordinary volume of research, but perhaps the most important catalyst has been the development of reliable and valid measures for assessing the core and associated symptoms of PTSD. The pivotal role of psychometrically sound assessment tools in PTSD research was underscored in the National Vietnam Veterans Readjustment Study (NVVRS; Kulka et al., 1990), the most extensive epidemiological study of PTSD ever conducted. According to Kulka et al., when the contract for the NVVRS was awarded in 1984, no validated measures of PTSD were yet available. Consequently, prior to conducting the main survey component of the study, the investigators were compelled to conduct a preliminary validation study in which various candidate measures were evaluated for their ability to distinguish PTSD cases from noncases. Several measures proved useful for this purpose and provided the means for achieving the primary research goal of estimating the prevalence of PTSD in combat veterans and matched controls.

Since then, the situation has changed substantially. The development and evaluation of standardized PTSD measures has become one of the most productive areas of research in the field of traumatic stress, and a wide variety of measures is now available, including questionnaires, structured interviews, and physiological protocols (for a recent review see Newman, Kaloupek, & Keane, 1996). Most of these measures enjoy at least some empirical support, and several, such as the Impact of Event Scale (IES; Horowitz, Wilner, & Alvarez, 1979), the Mississippi Scale for Combat-Related PTSD (Mississippi Scale; Keane, Caddell, & Taylor, 1988), the PK scale of the Minnesota Multiphasic Personality Inventory (MMPI) and MMPI-2 (Keane, Malloy, & Fairbank, 1984; Lyons & Keane, 1992), and the PTSD module of the Structured Clinical Interview for DSM-III-R (SCID; Spitzer, Williams, Gibbon, & First, 1990) have been examined extensively on diverse populations in a variety of settings. Given this abundance of instruments and a rapidly expanding empirical literature, researchers and clinicians can now choose instruments tailored to their particular assessment needs. Furthermore, they can increase their confidence in assessment decisions by relying on converging information obtained from multiple measures in an assessment battery, an approach that has been strongly advocated in the assessment of PTSD (Keane, Wolfe, & Taylor, 1987; Kulka et al., 1990).

The primary purpose of this chapter is to help readers become informed consumers of the burgeoning literature on the psychometric evaluation of PTSD assessment tools. It is intended primarily for those who wish to draw on this literature as a guide for selecting appropriate instruments for their particular assessment needs. Toward that end, we outline the key issues, principles, and techniques involved in developing and evaluating psychological assessment instruments, illustrating them with examples drawn from empirical work on a variety of PTSD measures. The chapter is divided into four sections. First, we describe a number of issues regarding scale construction and protocol development. Second, we discuss the concept of reliability from the perspective of classical test theory. Third, we discuss the concept of validity. Finally, we describe contemporary psychometric approaches, including generalizability theory, confirmatory factor analysis, and item response theory.

It is important to recognize that many of the concepts described in this chapter overlap with each other and that distinctions among them are not always clear-cut. For example, referring to the distinctions among content-related, criterion-related, and construct-related categories of validity, the *Standards for Educational and Psychological Testing* (1985) advises:

> The use of the category labels does not imply that there are distinct types of validity or that a specific validation strategy is best for each specific inference or test use. Rigorous distinctions between the categories are not possible. Evidence identified usually with the criterion-related or content-related categories, for example is relevant also to the construct-related category. (p. 9)

Under some circumstances the distinction between reliability and validity can also be blurred. For example, internal consistency statistics, usually interpreted as reflecting reliability across items on a scale, can also be seen as evidence of construct validity, in that a high degree of homogeneity among a set of items suggests they are measuring a single construct. Also, a study comparing PTSD diagnoses derived from a structured interview administered by lay interviewers versus experienced clinicians could be seen as a measure of test–retest reliability or as a measure of criterion-related validity, with the clinician's diagnosis as the standard against which the lay interviewer's ratings are compared. Despite the overlap, a broad distinction between reliability and validity and the further distinctions among the three categories of validity are meaningful and useful. In general, the two central concerns in the psychometric evaluation of an assessment instrument are the extent to which scores are free from measurement error (reliability) and the extent to which empirical evidence can be produced to demonstrate that the instrument measures what it purports to measure (validity).

Space constraints permit only the most general introduction to psychometric theory. Readers who plan to conduct research on existing PTSD measures or who are interested in developing new measures can find more in-depth discussions of the topics covered in this chapter in introductory texts such as Anastasi (1988), Crocker and Algina (1986), Cronbach (1990), and Suen (1990); in classic texts such as Gulliksen (1950), Lord and Novick (1968), and Nunnally (1978); and in the many works cited throughout the chapter. In addition, the *Standards for Educational and Psychological Testing*, cited earlier, is an essential reference for anyone involved in psychometric research, offering explicit, authoritative recommendations regarding the development, evaluation, and use of psychometric instruments. Finally, readers are encouraged to consult a recent special issue of the journal *Psychological Assessment* devoted to methods for increasing the psychometric integrity of psychological assessment instruments. This special issue makes an important contribution to the literature, with articles that (1) provide exhaustive coverage of topics such as content validity (Haynes, Richard, & Kubany, 1995); (2) discuss psychometric issues involved in assessment areas such as projectives (Weiner, 1995) and neuropsychology (Prigatano, Parsons, & Bortz, 1995); and (3) discuss psychometric issues involved in assessing special populations such as couples (Kashy & Snyder, 1995), substance abusers (Carroll, 1995), minorities (Okazaki & Sue, 1995), and the elderly (La Rue & Markee, 1995).

ISSUES IN SCALE CONSTRUCTION AND PROTOCOL DEVELOPMENT

Developing an effective psychological assessment instrument is a complex, painstaking, and iterative process. There is no single recipe that

can be followed, no set of necessary and sufficient steps that invariably produces the desired result. Rather, the creation of a useful scale requires expertise in the content area to be assessed, familiarity with the merits and shortcomings of various response formats, and a thorough understanding of the statistical concepts and procedures for establishing reliability and validity. It demands considerable ingenuity, careful trial and revision, and the accumulation of an extensive empirical database across a wide range of settings, populations, and assessment tasks.

Although different authors vary in terms of the number and sequence of the stages they identify in the test construction process, there is consensus about the major steps involved. For the purposes of this chapter we will consider five such stages: (1) specifying the purpose of the instrument, (2) defining the construct, (3) designing the instrument, (4) pilot testing and revising, and (5) establishing reliability and validity for the final version. In this section, we discuss the first four stages of the process. The remaining sections of the chapter are concerned with the conceptual issues and statistical techniques for accomplishing the last stage.

Specifying the Purpose of the Instrument

Before work on actually creating a new PTSD instrument can begin, it is essential to determine how it will be used. Identifying the major purposes for the instrument shapes the design of the instrument and enhances the likelihood that it will perform satisfactorily when it reaches final form. Most PTSD assessment instruments are designed to accomplish several assessment tasks, although some are better suited than others for certain applications. For example, questionnaires are simple to administer and yield information about overall severity of PTSD symptoms, but are not widely accepted in the role of diagnostic criterion or "gold standard." As in other areas of psychopathology, the most common "gold standard" in research on PTSD is a clinical diagnosis made on the basis of a structured interview. In a similar vein, physiological assessment procedures can provide direct evidence regarding reactivity to reminders of a trauma, but do not provide data regarding other PTSD symptoms.

Identifying the Assessment Tasks

Some of the most common reasons for developing a PTSD assessment tool are as follows:

1. *To obtain a continuous measure of the severity of the disorder.* Continuous measures provide relatively fine-grained information about PTSD symptom severity. Because they are capable of detecting subtle changes in symptom

status over time, they serve well as outcome measures. They are also useful in correlational analyses for testing hypotheses about PTSD in relationship to other constructs. All PTSD questionnaires and many structured interviews yield continuous scores. In addition, physiological measures yield continuous measures of reactivity in one or more response channels (e.g., heart rate, skin conductance, blood pressure), although it is more difficult to make the case that these are continuous measures of PTSD per se.

2. *To determine current and lifetime diagnostic status.* In many types of clinical research, including case-control, epidemiological, and treatment outcome studies, participants are classified as PTSD or non-PTSD. Accurate determination of PTSD "caseness" in these designs depends on the use of diagnostic instruments that yield present or absent decisions based on clear inclusion and exclusion criteria. PTSD diagnostic status is most often assessed with a structured interview. However, any continuous measure of PTSD, including questionnaires and physiological measures, can also serve as a dichotomous measure simply by selecting an appropriate cutoff score and dividing the sample into cases (scoring above the cutoff) and noncases (scoring below the cutoff).

3. *To assess distinct dimensions of symptom severity.* On most PTSD instruments, symptom severity is assessed as a single dimension, such as frequency or subjective distress. Items on the Clinician-Administered PTSD Scale (CAPS; Blake et al., 1990), however, were designed to assess the separate dimensions of frequency and intensity, with intensity ratings based on the additional dimensions of duration, subjective distress, and functional impairment.

4. *To assess associated features of PTSD.* A number of instruments, including the CAPS and questionnaires such as the Mississippi Scale, the Penn Inventory (Hammarberg, 1992), the Trauma Symptom Inventory (TSI; Briere, 1995), and the Los Angeles Symptom Checklist (LASC; King, King, Leskin, & Foy, 1995) include items to tap associated features of PTSD, including guilt, depression, substance abuse, and impairment of social and occupational functioning.

5. *To assess a range of stress response syndromes.* The DSM-IV symptoms of PTSD are widely acknowledged as encompassing some of the most significant sequelae of traumatic life events. Nonetheless, some clinical investigators, arguing that these symptoms do not adequately capture the full range of posttraumatic adjustment problems, have promoted the recognition of a spectrum of stress disorders, including acute stress disorder, PTSD, and complex PTSD. The TSI is an excellent example of an instrument designed specifically to assess the clinical phenomenology of all of these posttraumatic syndromes.

6. *To assess response bias.* As in other areas of psychopathology, the evaluation of response bias has been a central concern in the assessment of PTSD. Response bias may stem from efforts to minimize or exaggerate symptom severity, from carelessness, or from confusion or misinterpretation of test

items. Most PTSD measures are quite face valid, meaning the content they are intended to assess is obvious to respondents, making it easy to deliberately distort responses in a particular direction. For example, Lyons, Caddell, Pittman, Rawls, and Perrin (1994) found that the Mississippi Scale scores of combat veterans with PTSD were indistinguishable from those of three control groups without PTSD, who were instructed to respond as if they had the disorder. This suggests that it is relatively easy for respondents to shape their answers to create a desired impression, and in the absence of additional information regarding the validity of responses, such distortion would be impossible to detect. Currently, few PTSD measures include items to assess response bias. The PK scale, when scored from the full MMPI or MMPI-2, can be interpreted in the context of the validity scale profile. Also, the TSI contains three scales to evaluate different aspects of response bias. The CAPS, the only interview to address this issue, assesses response validity at the item level and as a global rating that takes into account the entire interview. Unfortunately, these ratings, especially at the item level, are very difficult to make reliably (Weathers & Litz, 1994).

7. *To assess PTSD in the context of existing instruments.* PTSD scales have been developed from existing measures of psychopathology, including the PK scale and the crime-related (CR-PTSD; Saunders, Arata, & Kilpatrick, 1990) and war-zone-related (WZ-PTSD; Weathers et al., 1996) PTSD scales of the Symptom Checklist 90—Revised (SCL-90-R; Derogatis, 1983). Using these measures brings the additional benefits of the parent instrument, such as the assessment of response validity and the assessment of comorbid problems such as depression, anxiety, and substance abuse. The availability of these measures also permits the assessment of PTSD in archival data sets collected before the existence of instruments developed specifically for PTSD.

8. *To quantify some directly observable PTSD-relevant behavior.* This is the primary goal both in the development of physiological assessment procedures (e.g., Malloy, Fairbank, & Keane, 1983; Pitman, Orr, Forgue, de Jong, & Claiborn, 1987) and in the development of protocols assessing other aspects of PTSD, such as emotional numbing (Litz, 1992) and intimacy deficits (Weathers et al., 1995).

Identifying the Target Populations

Another critical aspect of specifying the purpose of a PTSD assessment tool is to identify the populations in which it will be used. Much of the early research on PTSD assessment instruments was conducted on combat veterans, and both the PK scale and the Mississippi Scale, were originally developed specifically for this population. For the Mississippi Scale, this choice of populations dictated item content, with many items referring directly to experiences in the military. The use of veterans also undoubtedly influenced the MMPI items included in the PK scale. Although there is some evidence that the PK scale is useful in civilian trauma populations (e.g., Koretzky &

Peck, 1990), indicating some generalizability of content, other items likely would have emerged had the scale been developed in a population of rape victims, for example.

The importance of specifying the target population is illustrated by recent efforts to develop PTSD scales on the SCL-90-R (Derogatis, 1983). In developing the CR-PTSD scale, Saunders et al. (1990) identified 28 SCL-90-R items that differentiated crime victims with and without PTSD. Weathers et al. (1996) employed the same methodology in a sample of combat veterans, but found that only 11 of the 25 items on their WZ-PTSD scale overlapped with items on the CR-PTSD scale. The influence of the target population can sometimes be more qualitative and subtle, and requires experience with a given population and extensive pilot testing. For example, in our work with male combat veterans, we have found that in evaluating subjective distress, even slight changes in the phrasing of probe questions can evoke very different responses. Specifically, descriptors such as "upset" and "afraid" (How *upset [afraid]* did you feel?) tend to elicit lower levels of endorsement relative to "distress or discomfort" or "bothered" (How much *distress* or *discomfort* did you feel? How much did it *bother* you?).

Increasingly, test developers have created instruments that would measure PTSD in any population and have begun to validate and norm their instruments in diverse groups of trauma survivors. Recent examples of such measures include the Penn Inventory, the civilian version of the Mississippi Scale (Keane et al., 1988; Vreven, Gudanowski, King, & King, 1995), and the TSI. This approach should promote standardization of PTSD assessment and lead to an evaluation of the common aspects of PTSD resulting from any type of traumatic life event. Two population specifications that will likely remain important for instrument construction are the distinction between children and adults, and the distinction between victims of single or circumscribed traumatic events and victims of chronic interpersonal trauma. Each of these distinctions has implications for the format and content of test items.

Defining the Construct

Defining the construct to be measured and determining appropriate item content is another crucial step in the development of a PTSD assessment tool. Inadequate specification of the construct could result in the exclusion of items measuring core features of PTSD or the inclusion of items tapping irrelevant content. Either of these undesired outcomes could result in misleading correlations between PTSD and other constructs, or in diagnostic inaccuracies that could, in turn, affect prevalence estimates or formation of distinct groups for case control research. Unfortunately, PTSD has proven to be a difficult construct to define, and there is continued debate regarding the appropriate boundaries of the syndrome and its relationship to other disorders and other hypothesized sequelae of traumatic life events.

The formal diagnostic criteria for PTSD, first introduced in the DSM-III in 1980, constitute the most widely accepted and frequently invoked operational definition of the disorder. The advantage of this definition is that it is a consensus standard reflecting the most current clinically and empirically informed conceptualization of PTSD. The disadvantage is that the diagnostic criteria, and thus the construct, have evolved considerably as more information about the disorder has become available. Some notable changes from the DSM-III to the DSM-IV (American Psychiatric Association, 1994) include (1) redefining the stressor criterion; (2) adding a distinct hyperarousal symptom cluster; (3) combining numbing and avoidance symptoms into the same cluster; (4) dividing cued "symptom intensification" into cued physiological arousal and psychological distress; (5) adding avoidance of thoughts and feelings; (6) dropping guilt and nonspecific memory impairment, but adding memory impairment related to the trauma; (7) adding "sense of a foreshortened future," based primarily on research with traumatized children, and qualifying several criteria to reflect alternative symptom expression in children; and (8) adding the requirement that the syndrome cause significant distress or impairment in social or occupational functioning.

These changes in the diagnostic criteria reflect the difficulties in identifying and articulating the core symptoms of PTSD. Undoubtedly these criteria will continue to evolve, posing a moving target for developers of PTSD assessment tools. In addition, several other factors extend and blur the boundaries of PTSD, further complicating efforts to derive an explicit operational definition. First, PTSD has consistently been found to be associated with high rates of comorbid syndromes, including depression, anxiety, substance abuse, and dissociation (e.g., Kulka et al., 1990; Orsillo et al., 1996). The reasons for this comorbidity are not clear, but it has prompted some investigators to challenge the assumption that PTSD is a distinct diagnostic entity. Even those who view PTSD as a unique disorder have raised questions regarding its appropriate placement in the current psychiatric taxonomy. Davidson and Foa (1991), for example, recently reviewed clinical and laboratory findings variously supporting the classification of PTSD as an anxiety disorder, as a dissociative disorder, or as part of a new etiologically based category of stress disorders.

Second, some clinical researchers have contended that the current diagnostic criteria for PTSD do not adequately represent the full range of posttraumatic symptomatology, especially in victims of chronic interpersonal trauma such as physical and sexual abuse or spouse battering. For example, arguing for a broader conceptualization of posttraumatic sequelae, Herman (1992) has promoted the recognition of a diagnostic category that she labels "complex PTSD," which includes symptom clusters such as affective dysregulation, dissociation, and self-destructive behaviors. Third, emphasizing the impact of traumatic events on core beliefs about the self and the world, constructivist models of traumatic stress (e.g., Janoff-Bulman, 1992; McCann & Pearlman, 1990) have expanded the domain of posttraumatic problems be-

yond the realm of PTSD symptoms per se. All of these issues reflect the controversy and ambiguity currently surrounding the PTSD construct and speak to the practical difficulty test developers face in deriving an unambiguous definition of the disorder as the basis of a useful assessment tool.

Most structured PTSD interviews, such as the PTSD module of the SCID, are based on the DSM-III-R (and now DSM-IV) conceptualization of PTSD and follow the diagnostic criteria closely. The CAPS also follows the DSM criteria, but includes additional questions assessing social and occupational impairment and associated features such as guilt and dissociation. Among PTSD questionnaires, however, there is considerably more variability in how the construct is defined, and thus greater diversity in item content. The IES, developed prior to the availability of formal diagnostic criteria, was based on clinical observations of individuals suffering from what Horowitz et al. (1979) termed "stress response syndromes." Items were written to assess what were then viewed as the two primary symptom clusters developing in response to stressful life events, intrusions and avoidance, so the IES does not assess hyperarousal symptoms.

The Mississippi Scale, developed for use in combat veterans, was based primarily on the DSM-III diagnostic criteria for PTSD, but Keane et al. (1988) included items tapping associated features of combat-related PTSD, including depression, substance abuse, and impairment in social and occupational functioning. The Penn Inventory was based on DSM-III and DSM-III-R criteria, but also includes items assessing identity confusion, spirituality, and goal-directed behavior. The TSI, easily the most comprehensive of all current questionnaires, was designed to encompass a broad spectrum of posttraumatic symptomatology, including the core symptoms of PTSD as well as a variety of symptoms associated with complex PTSD and acute stress disorder. Finally, like the structured interviews, some PTSD questionnaires such as the PTSD Symptom Scale (PSS; Foa, Riggs, Dancu, & Rothbaum, 1993) and the PTSD Checklist (PCL; Weathers, Litz, Herman, Huska, & Keane, 1993) are based directly on the DSM diagnostic criteria.

The PTSD construct undoubtedly will continue to evolve, as a result, we hope, of accumulation of scientific evidence supporting the inclusion or exclusion of various phenomena. New measures will emerge and existing measures will require periodic revision. As advances are made, it will remain incumbent on PTSD test developers to explicate and defend precise definitions of the construct as they conceptualize it. It should be noted that clear definitions of the construct are less central in the development of empirically derived measures, such as the PK scale and the PTSD scales of the SCL-90-R, in that items on these scales are selected only for their ability to discriminate those with and without PTSD. Nonetheless, the significance of the definition of the construct is implicit in terms of how PTSD and non-PTSD groups are defined, since some PTSD measure, with its attendant definition, must be employed to determine caseness.

Designing the Instrument

After specifying the purpose and defining the construct to be measured, the next stage in developing a PTSD assessment tool is designing the new instrument. The steps involved in designing a scale vary depending on whether the scale is rationally derived or empirically derived. As noted earlier, empirically derived PTSD instruments consist of items on existing scales that statistically differentiate PTSD cases and noncases. In contrast, rationally derived instruments consist of new items based on clinical observation and theoretical conceptualizations of the construct being measured. Since it involves the creation of new items, developing rationally derived instruments is a much more elaborate process. In general, designing the instrument entails five tasks.

Determining the Length of the Scale

The first task in designing a rationally derived measure is to determine the overall length of the scale and approximately how many items will be devoted to each aspect of the construct. Also, test developers must decide whether response bias will be addressed, either by reversing some proportion of items or by including additional items specifically for that purpose. Most rationally derived PTSD scales contain fewer than 40 items and do not include items to evaluate response bias. For example, the Mississippi Scale consists of 35 items and addresses the issue of response bias only through the inclusion of reversed items (e.g., "I still enjoy doing many things I used to enjoy"). A notable exception is the TSI, a 100-item instrument yielding 10 clinical scales and 3 validity scales. For empirically derived scales, this task consists simply of determining the number of items to retain. In developing the PK scale, Keane et al. (1984) identified 49 MMPI items that differentiated veterans with and without PTSD at $p < .001$, and retained these items on the final scale. This task can be more difficult, however. In developing the WZ-PTSD scale of the SCL-90-R, Weathers et al. (1996) found significant differences between PTSD and non-PTSD groups on nearly every item on the SCL-90-R. In order to reduce the number of items on the final scale, they decided to employ a more stringent statistical criterion, retaining 25 items that offered the greatest discrimination between the groups.

Selecting an Item Format

The second task is to select an item format. For rationally derived questionnaires, this involves designing both the item-stem and response-option formats. Most PTSD questionnaires employ a Likert-type format, with items consisting of brief statements assessing PTSD symptoms and a response continuum (typically a 5-point scale) indicating level of endorsement or agree-

ment, frequency or severity of symptoms, or degree of subjective distress. Most questionnaires employ a single response option format for all items. The IES, for example, assesses symptom frequency on a 4-point scale ranging from "not at all" to "often." Similarly, the PCL assesses subjective distress on a 5-point scale ranging from "not at all" to "extremely." In contrast, the Mississippi Scale contains items assessing symptom frequency ("If something happens that reminds me of the military, I become very distressed and upset," with five response options ranging from "never" to "very frequently"), as well as items assessing symptom endorsement ("No one understands how I feel, not even my family," with response options ranging from "not at all true" to "extremely true"). A notable exception to the Likert-type format is the Penn Inventory, which is modeled after the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and consists of items comprising four graded statements reflecting increasing levels of psychopathology.

For PTSD interviews, the typical item format consists of one or more standard prompt questions and a rating scale to evaluate symptom severity. The SCID-PTSD module, for example, consists of single prompt questions for each symptom and a three severity rating options, "absent," "subthreshold," and "present," although these ratings are typically treated as dichotomous (present–absent) judgments. Several recently developed interviews, including the PTSD Interview (PTSD-I; Watson, Juba, Manifold, Kucala, & Anderson, 1991) and the Structured Interview for PTSD (SI-PTSD; Davidson, Smith, & Kudler, 1989), provide a broader continuum of rating options and yield both continuous and dichotomous scores, both for individual items and for the disorder. Employing the most elaborate item format of any PTSD interview, the CAPS assesses the frequency and intensity of symptoms separately by means of initial and follow-up prompt questions and behaviorally anchored 5-point rating scales.

Developing an Item Pool

For rationally derived PTSD scales, the next task is to create items to assess the various aspects of the construct. This task requires considerable expertise in the clinical phenomenology of the disorder and is usually based on clinical observation and descriptions of trauma-related symptoms in the literature. For questionnaires, test developers typically create many more items than they plan to incorporate into the final form of the instrument, eliminating items through pilot testing and review. Apart from their correspondence with key aspects of the construct, what are some of the desirable characteristics of items on questionnaire measures of psychopathology? Holden and Fekken (1990) recently addressed this question in an intriguing investigation. Drawing on extensive analysis of items on the Basic Personality Inventory (BPI; Jackson, 1976) they suggested that good items (1) are free from negatives and absolutes, (2) can be answered quickly, (3) include undisguised

contents and inquire about pathological or unusual content when appropriate, (4) emphasize general behavior tendencies, and (5) contain comparisons to other people or include statements by other people.

Reviewing and Revising Items

Once the initial item pool is created, items should be reviewed by experts in PTSD. Items should be evaluated in terms of their match to the symptoms being assessed, and in terms of readability, degree of ambiguity, and redundancy with other items. Feedback from the review process serves as the basis for revising or eliminating items from the pool. For example, in developing the Mississippi Scale, Keane et al. (1988) began with a pool of 200 items, which they eventually reduced to the 35 items on the final scale through a process of expert review. Similarly, in developing the TSI, Briere (1995) began with a pool 182 items, reducing it to the 100 items on the final scale through a combination of expert review and preliminary item analyses on a sample of respondents.

Specifying the Protocol

The final task in designing a PTSD assessment tool is to specify the protocol for administering and scoring the instrument. The protocol should include instructions for the clinician, specifying the target populations, standard testing conditions, procedures for administration, explicit scoring rules, and guidelines for interpreting scores. It should also include instructions to respondents, explaining the format and describing or illustrating appropriate responses. For questionnaires, the protocol is typically straightforward, with administration involving only a few brief instructions to respondents, and scoring consisting of simply summing the items to obtain a total score. For interviews the protocol can be more elaborate and less structured. For example, interviewers may need to clarify prompt questions ad lib, or they may need to determine whether to skip out of a section, or to inquire about lifetime symptom status. Scoring interviews may also be more complicated, particularly if an interview yields both continuous and dichotomous diagnostic scores.

Pilot Testing and Revising the Instrument

After the initial item pool has been created and reviewed, and the protocol has been specified, the instrument should be pilot tested on a sample of respondents from the target population. The purpose of such pilot testing is to conduct preliminary statistical analyses of individual items, and to obtain additional qualitative feedback about item content. The techniques for this process are described here. This information is then used to further revise items or to eliminate additional items as the scale nears its final form.

Once the final form has been reached, test developers set about exploring its reliability and validity.

ESTABLISHING RELIABILITY AND VALIDITY

Reliability

In general, reliability refers to the consistency of test scores over repeated measurements. If a test is reliable, it means that respondents achieve the same or nearly the same score each time they are evaluated. However, all psychological assessment procedures are unreliable to some extent, meaning they entail some degree of measurement error, so observed scores for a group of respondents are never exactly reproducible upon retesting. The classical test theory approach to reliability (Crocker & Algina, 1986; Gulliksen, 1950; Lord & Novick, 1968) is based on a simple formula depicting this point:

$$X = T + E$$

From this perspective, an observed test score (X) is viewed as the sum of two components, a respondent's true score (T), and an error component (E). True scores have been variously defined and interpreted but, in general, reflect a respondent's actual standing on the attribute being measured. Measurement errors may be either systematic or random. Systematic errors, such as a tendency to use the extreme ends of a rating scale to exaggerate or minimize symptom severity, are relatively constant across repeated measurements. Although they reflect inaccuracies, systematic errors do not contribute to inconsistency in observed scores and thus do not reduce reliability. In contrast, random errors fluctuate across repeated measurements and thus play an important role in the reliability of an assessment instrument. Random errors can arise from a number of sources, including lapses in concentration, fluctuations in mood, carelessness, adoption of a random response set, variability in interpretation of ambiguously worded items, or variability in responses due to characteristics of the examiner (e.g., age, gender, ethnicity).

Classical test theory makes three key assumptions about random errors. Across a large number of observations, random errors are assumed to have a mean of zero ($M_E = 0$), to be uncorrelated with true scores ($r_{TE} = 0$), and to be uncorrelated with errors on other testing occasions ($r_{E1E2} = 0$). It can then be shown that

$$s^2_X = s^2_T + s^2_E$$

This formula indicates that observed score variance consists of true score variance and error variance. For reliable tests, error variance is minimal, and thus most of the variability in observed scores is attributable to genuine differences among respondents.

Classical test theory defines the "reliability coefficient" as the correlation between observed scores on parallel tests. Two tests are considered parallel if they yield the same true scores and have equal error variances. Drawing on these definitions and the assumptions about random errors, it can be shown that

$$r_{X1X2} = s^2_T / s^2_X$$

That is, the "reliability coefficient," defined as the correlation between two parallel tests, is equivalent to the ratio of true score variance to observed score variance. Thus, the correlation coefficient can range from 0 (if observed scores are completely random and consist only of error variance) to 1 (if observed scores are completely free from measurement error and consist only of true score variance).

Several procedures have been developed for obtaining estimates of reliability (Crocker & Algina, 1986). Some procedures involve administering two tests. The alternate forms method requires constructing two comparable forms of a test and administering both within a very short time span. The reliability coefficient obtained by correlating scores on the two forms is called the "coefficient of equivalence." The test–retest method involves administering the same instrument twice, with some reasonable time period between administrations, and the resulting reliability coefficient is called the "coefficient of stability," or simply the "test–retest reliability coefficient."

Other procedures permit reliability estimates from a single test administration. The split-half method involves dividing the items on an instrument in half, correlating the scores from the two half-tests, then using the Spearman–Brown formula to calculate the projected reliability for the whole test. The general form of the Spearman–Brown formula, which can be used to estimate reliability following any change in the length of a test, is

$$r_{SB} = kr_{XX} / [1 + (k - 1)r_{XX}]$$

where r_{SB} is the reliability of the altered test, r_{XX} is the reliability of the original test, and k is a factor indicating how much longer ($k > 1$) or shorter ($k < 1$) the new test will be. So, if the correlation between the two halves of a test were .60, the projected reliability coefficient for the full test would be $2(.60)/[1 + (.60)] = .75$. The main problem with the split-half approach is that there are many ways to divide the test items in half, and each division could yield a different reliability coefficient. For that reason, test developers rely on measures of internal consistency instead, most commonly Cronbach's alpha (Cronbach, 1951). Coefficient alpha reflects item homogeneity, or the degree to which items are intercorrelated, and can be interpreted as the mean of all possible split-half reliability coefficients.

Another useful index of measurement error is the standard error of measurement, which can be used to construct confidence intervals around

observed scores for individual respondents. The formula for the standard error of measurement is

$$s_E = s_X \sqrt{1 - r_{XX}}$$

where s_X is the standard deviation of observed scores, and r_{XX} is the reliability coefficient. The general form of the 95% confidence interval (CI) for an observed score is $X \pm 1.96\, s_E$. So, for example, if a respondent obtained a score of 50 on a test with $s_X = 10$ and $r_{XX} = .84$, then $s_E = 4$ and the 95% CI would be (with rounding) 50 ± 8. Thus, there is a 95% probability that the interval from 42–58 contains the respondent's true score.

It is important that all reliability coefficients be reported and interpreted with reference to the specific procedures used to estimate them, since each procedure takes into account different sources of error. The test–retest method takes into account variability due to changes in respondents and changes in the testing conditions, but not variability due to items. In contrast, internal consistency and split-half methods primarily take into account variability due to item content. The alternate forms method also takes into account variability due to items, but if the interval between administrations of the two forms is sufficiently long, it may also be affected by the same sources of variability as the test–retest method. Generalizability theory (to be described) makes these design considerations explicit and provides a much more flexible framework for conceptualizing and evaluating potential sources of measurement error.

The measures of reliability described thus far are appropriate for assessment measures that yield continuous scores. Different techniques for estimating reliability are needed when the scores of interest are dichotomous, such as present–absent diagnostic decisions. One approach is to calculate the percentage of agreement between two raters by counting the number of times both agree that the diagnosis is either present or absent and dividing by the total number of respondents. The problem with this approach is that it fails to take into account the fact that a certain amount of agreement would be expected by chance. The kappa statistic (Cohen, 1960) was developed to overcome this limitation. The formula for kappa is

$$\kappa = (P_o - P_c) / (1 - P_c)$$

where P_o is the proportion of observed agreement, and P_c is the proportion of agreement due to chance. As this formula indicates, kappa provides an index of chance-corrected agreement. A kappa of 0 does not mean that there is no agreement, only that the agreement does not exceed the level expected by chance. A kappa of 1 indicates perfect agreement. There is some disagreement among investigators as to how large a kappa is needed to reflect a satisfactory level of reliability. Fleiss (1981) addressed this issue, describing kappas from .40 to .60 as fair, kappas from .60 to .75 as good, and kappas above .75 as excellent.

In the evaluation of PTSD measures, the two most commonly reported forms of reliability are test–retest reliability and coefficient alpha. A crucial consideration in estimating test–retest reliability is the duration of the interval between administrations of an instrument. The interval must be long enough to reduce memory and practice effects, but brief enough so that scores are not greatly affected by genuine changes in PTSD symptom severity. In most instances, investigators have selected intervals ranging from a few days to a week. Another commonly employed strategy, useful for identifying poor items on a scale, is to examine the correlations of individual items with the total score, and to examine the changes in alpha that result from removing one item at a time.

In the original report on the Mississippi Scale, for example, Keane et al. (1988) found a test–retest reliability of .97 at a 1-week interval and an alpha of .94. Examining individual items, they found that item-total correlations ranged from .23 to .73 with a mean of .58. Similarly, for the Penn Inventory, Hammarberg (1992) found a test–retest reliability of .96 at approximately a 5-day interval, an alpha of .94, and item-total correlations ranging from .43 to .90, with a mean of .75. Excellent reliability has also been found for PTSD interviews. For example, Weathers et al. (1992) administered the CAPS twice with independent clinicians at a 2- to 3-day interval. Test–retest reliability for CAPS total severity scores ranged from .90 to .98 across three different rater pairs, and alphas ranged from .85 to .87 for the three PTSD symptom clusters, with an alpha of .94 for all 17 symptoms. Using the CAPS as a diagnostic instrument, they found a kappa of .89 for the optimal diagnostic scoring rule.

Similar findings have been reported for many other PTSD measures, indicating that it is possible to reliably assess PTSD symptom severity and diagnostic status. One unfortunate aspect of the growing literature on the reliability of PTSD assessment tools is that investigators almost never report standard errors of measurement. This is a regrettable lapse, since these would provide a useful tool for interpreting individual scores, particularly when cutoff scores are employed for diagnostic decision making. If a respondent scores just above or just below a suggested cutoff, it would be helpful to employ a confidence interval in determining caseness. Finally, although reliability is important, validity is of even greater concern. The next section describes various methods for evaluating different types of validity and illustrates their application to various PTSD instruments.

Validity

"Validity" is a general term referring to the scope and quality of evidence supporting the inferences, interpretations, classifications, decisions, or predictions made on the basis of test scores. According to the *Standards for Educational and Psychological Testing* (1985):

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (p. 9)

Three different types of validity are usually considered: content validity, criterion-related validity, and construct validity.

Content Validity

"Content validity" refers to the extent to which items or stimuli on a psychological assessment instrument measure key aspects of the construct being evaluated. As discussed earlier, consideration of item content is primarily relevant for rationally derived instruments and is less of a consideration for empirically derived instruments. Haynes et al. (1995) recently presented a detailed conceptual analysis of content validity and provided an exhaustive set of practical guidelines for establishing content validity for a wide variety of assessment instruments and tasks. They describe more fully many of the procedures outlined earlier in the chapter in the sections on specifying the purpose of the instrument, defining the construct, and designing the instrument. Identifying more than a dozen potentially relevant steps in the content validation process, they emphasize the iterative, sequential nature of developing a quality assessment tool. Furthermore, they highlight the indispensable role of expert judgment in generating and revising items, and in determining the extent to which the item pool provides sufficient coverage for all important facets of the construct to be measured.

Criterion-Related Validity

"Criterion-related validity" refers to the ability of a measure to accurately predict some outcome variable of interest. When the predictor and the criterion are assessed simultaneously, this is referred to as "concurrent validity," and when the criterion is assessed at some point in the future, this is referred to as "predictive validity." In research on PTSD assessment tools, the most common application of criterion-related validity is in the evaluation of diagnostic utility, or the ability of a measure to predict diagnostic status. A central concern in establishing criterion-related validity is the question of how best to define the criterion. Although other procedures for establishing caseness have been proposed (e.g., Kulka et al., 1990; Spitzer, 1983), the criterion or so-called "gold standard" most often employed in PTSD research is a diagnosis based on a structured clinical interview.

A typical investigation of the diagnostic utility of a PTSD questionnaire, for example, involves (1) administering the questionnaire and a structured diagnostic interview to a sample of respondents, (2) diagnosing each respondent as PTSD or non-PTSD on the basis of the interview, (3) selecting a cutoff score on the questionnaire and dichotomizing the sample into test positives (scoring at or above the cutoff) or test negatives (scoring below the cutoff) on the basis of the structured interview, and (4) constructing a 2 × 2 table to compare the questionnaire to the diagnosis. As shown in Figure 4.1, four outcomes are possible in this procedure, two involving agreement between the test and the diagnosis, and two involving disagreement or classification errors. In terms of agreement, respondents with a positive diagnosis and a positive test are called "true positives," and those with a negative diagnosis and a negative test are called "true negatives." In terms of disagreement or errors, respondents with a negative diagnosis but a positive test are called "false positives," and those with a positive diagnosis but a negative test are called "false negatives."

Figure 4.1 also provides definitions of various indices of diagnostic utility, including (1) sensitivity (the probability of a positive test, given a posi-

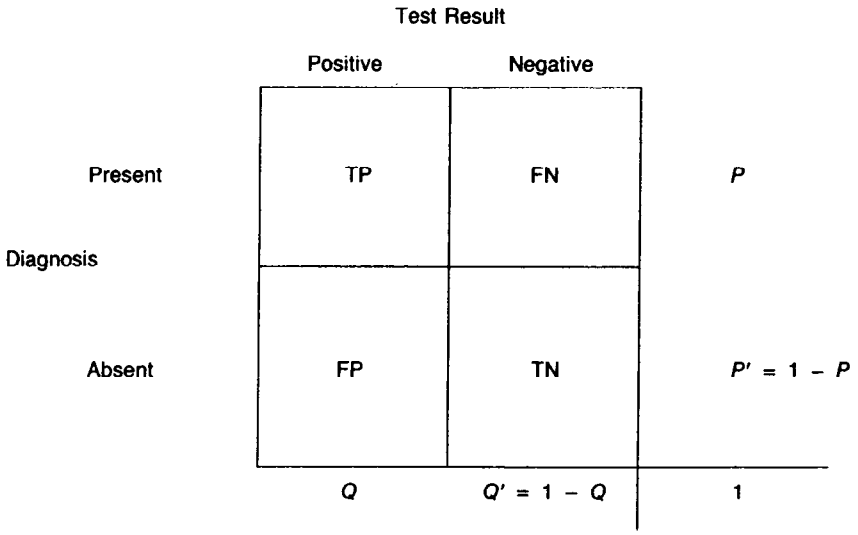


FIGURE 4.1. Definition of key terms in evaluating the diagnostic utility of a PTSD instrument. Based on Kessel and Zimmerman (1993) and Kraemer (1992).

tive diagnosis); (2) specificity (the probability of a negative test, given a negative diagnosis); (3) positive predictive power (the probability of a positive diagnosis, given a positive test); (4) negative predictive power (the probability of a negative diagnosis, given a negative test); and (5) efficiency or overall correct classification (the probability that the test and the diagnosis agree). The diagnostic utility for all possible cutoff scores on a questionnaire can be evaluated by constructing separate 2 × 2 tables and calculating these various probabilities.

Across the range of possible cutoff scores, there is a trade-off between sensitivity and specificity, with lenient scores having higher sensitivity but lower specificity, and stringent scores having lower sensitivity but higher specificity. This relationship can be depicted by plotting sensitivity against specificity for each possible cutoff score, which results in a Receiver Operating Characteristic (ROC) curve (see Kraemer, 1992). Figure 4.2 compares the ROC curves for three PTSD instruments. The ROC curve for the Mississippi Scale extends nearest the ideal test point in the upper-right corner, indicating that across a range of cutoffs, it is the best predictor of a PTSD diagnosis.

Although measures of test performance such as sensitivity, specificity, and efficiency, depict the relationship between test and diagnosis, Kraemer

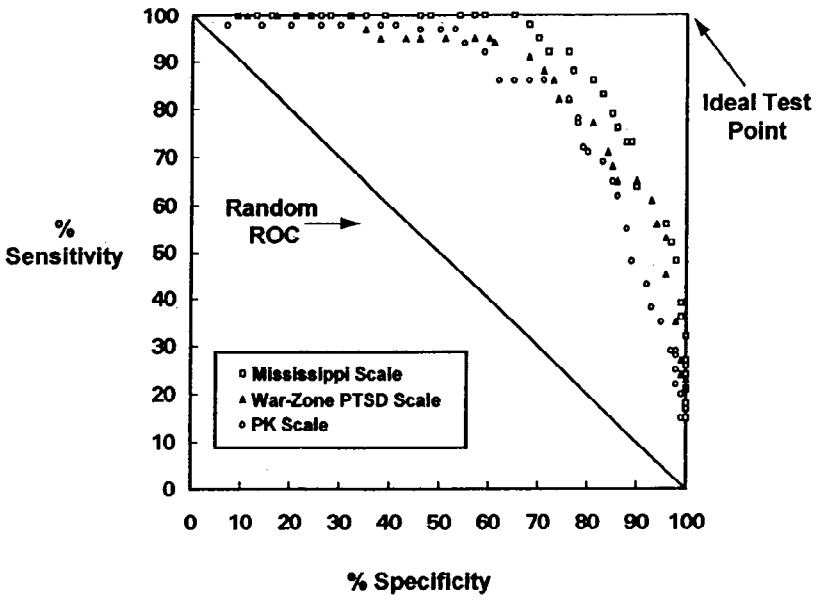


FIGURE 4.2. Receiver Operating Characteristic (ROC) curves for the Mississippi Scale, PK scale, and WZ-PTSD scale. From Weathers et al. (1996). Copyright 1996 by Plenum Press. Reprinted by permission.

(1992) has shown that they are ambiguous indicators of diagnostic utility, because they are uncalibrated and do not take into account chance agreement between test and diagnosis. She proposed the use of weighted kappa coefficients as indicators of the quality of sensitivity [$\kappa(1)$], specificity [$\kappa(0)$], and efficiency [$\kappa(.5)$]. The measures of test quality are calibrated such that a value of 0 represents chance agreement, and a value of 1 represents perfect agreement.

Plotting the quality of sensitivity against the quality of specificity results in the Quality Receiver Operating Characteristic (QROC) curve, which permits ready identification of the optimally sensitive, specific, and efficient cutoffs. Figure 4.3 compares the QROC curves for the same three PTSD instruments shown in Figure 4.2. The QROC curves are a one-to-one remapping of the ROC curves that permits straightforward identification of the optimally sensitive, specific, and efficient cutoffs. These various cutoff scores are useful for different assessment tasks. Optimally sensitive cutoffs are useful for screening, optimally specific cutoffs are useful for making definitive diagnoses, and optimally efficient cutoffs, which are the cutoffs most often reported in the literature, are useful for differential diagnosis. Figure 4.3 shows that the Mississippi Scale outperforms the other instruments with respect to the quality of sensitivity, specificity, and efficiency.

Construct Validity

Construct validity is the most encompassing category of validity, and it can be argued that all types of validity evidence are relevant for construct validity (Cronbach & Meehl, 1955; *Standards for Educational and Psychological Testing*, 1985). "Construct validity" refers to the extent to which a pattern of evidence exists supporting the interpretation of a test as a measure of some underlying attribute. The accumulation of relevant evidence must be guided by an explicit theory in which the construct to be measured is defined and its hypothesized relationships with other constructs are made explicit.

Test developers may draw on a number of different sources of evidence in the construct validation of an assessment instrument. An important source of evidence is the pattern of correlations between the instrument and measures of other constructs. Ideally, the instrument should correlate strongly with other measures of the same construct (convergent validity) and should correlate weakly with measures of other constructs (discriminant validity). Campbell and Fiske (1959) provided an elegant framework for evaluating convergent and discriminant validity, which they referred to as the multitrait–mulitmethod matrix. In this approach, each of several distinct constructs is measured by each of several methods, and the resulting correlation matrix is examined to determine if correlations between different measures of the same construct exceed those between measures of different constructs. For example, in an effort to evaluate the construct validity of the CAPS, Weathers et al. (1992) assessed each of four constructs (PTSD, depression, anxiety, and

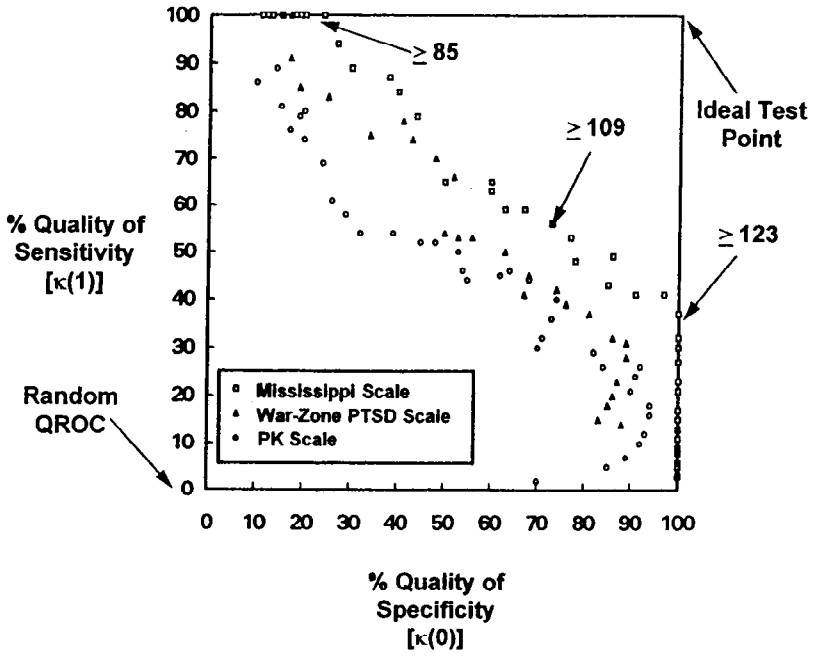


FIGURE 4.3. Quality Receiver Operating Characteristic (QROC) curves for the Mississippi Scale, PK scale, and WZ-PTSD scale. From Weathers et al. (1996). Copyright 1996 by Plenum Press. Reprinted by permission.

antisocial personality disorder) using three different methods (structured interview, dedicated questionnaire, and MMPI scale). They found that the resulting pattern of correlations generally conformed to predictions. The CAPS correlated strongly with other measures of PTSD, correlated moderately with measures of depression and anxiety, and correlated weakly with measures of antisocial personality.

Other sources of evidence for construct validity include (1) evidence for content validity, especially expert judgment regarding the appropriateness of items for the construct being measured; (2) the internal consistency of items, which may be taken as evidence that the instrument is measuring a single construct; (3) factor analysis, which may reveal theoretically meaningful dimensions underlying test scores (to be discussed); (4) differences in test scores among groups hypothesized to vary in the underlying construct; and (5) changes in test scores as a result of treatment or some other intervention hypothesized to directly influence respondents' standing on the construct.

CONTEMPORARY APPROACHES

Generalizability Theory

Generalizability theory (G theory) was developed by Cronbach and his colleagues (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Brennan, 1983; Shavelson & Webb, 1991) in order to address the ambiguities and limitations of the classical-test-theory approach to reliability. Like classical test theory, G theory is concerned with the replicability of scores across repeated measurements, but G theory replaces the classical-test-theory concept of a true score with the concept of a universe score. As Cronbach et al. (1972) explain:

> A behavioral measurement is a sample from the collection of measurements that might have been made, and interest attaches to the obtained score only because it is representative of the whole collection or *universe*. If the decision maker could, he would measure the person exhaustively and take the average over all the measurements. Educators and psychologists have traditionally referred to the average reached via exhaustive measurement as "the true score" for the person. We speak instead of a *universe score*. This emphasizes that the investigator is making an inference from a sample of observed data, and also that there is more than one universe to which he might generalize. . . . "The universe score is estimated to be 75" is without meaning until we answer the question, "Which universe?" This ambiguity is concealed in the statement "The estimated true score is 75," for no one thinks to inquire, "Which truth?" (pp. 18–19)

Thus, from a G-theory perspective, investigators must explicitly describe what Cronbach et al. refer to as the universe of admissible observations, or the entire set of observations that they would be willing to accept as interchangeable for a given assessment task. They can then determine the extent to which respondents' scores from a single observation generalize to their universe scores.

G theory is based on an analysis of variance (ANOVA) framework. Different conditions or facets of measurement, which represent potential sources of error, are employed as factors in an experimental design, and the proportion of variance attributable to each facet is estimated through the techniques of ANOVA. In classical test theory, only one facet at a time can be considered. Test–retest reliability, for example, is concerned primarily with the facet of occasions. Similarly, internal consistency is concerned primarily with the facet of items. A significant advantage of G theory is that multiple facets of observation can be represented in the same design, and there are few restrictions on the type of facets that can be examined. This permits much more elaborate experimental designs and allows for the estimation of variance due to the interaction of two or more facets.

G theory distinguishes between G studies and D studies. In G (general-

izability) studies, multiple facets are included in an experimental design in order to identify as many significant sources of variability in test scores as possible. Data from a comprehensive G study can then be put to use in designing a D (decision) study, in which some decision or conclusion about respondents will be reached on the basis of test scores. Although the emphasis in G theory is on the estimation of variance components attributable to respondents or to one or more facets of observation, G coefficients, analogous to reliability coefficients in classical test theory, can be calculated for a variety of different D studies.

G theory also distinguishes between relative and absolute decisions, and different G coefficients are calculated for each type of decision. "Relative decisions" refer to situations in which test scores are used to indicate a respondent's relative standing in a distribution, such as when the correlation between two different measures is explored. "Absolute decisions" refer to situations in which test scores are used to indicate a respondent's status with respect to a fixed standard, such as on a driver's license examination. In contrast, classical test theory reliability coefficients only apply to relative decisions.

To date, investigators have not applied G theory in developing PTSD assessment tools. This is unfortunate, given the power and flexibility of this approach. In the following example we illustrate the application of G theory to the development of a physiological assessment protocol measuring physiological and psychological reactivity to slides and sounds depicting combat scenes. The data for this example are fictitious, but are based on data taken from the physiological protocol in use at our PTSD clinic. Table 4.1 presents the results of a two-facet G study. In this design, 20 respondents observed six slides on each of two occasions. As shown in Table 4.1, nearly half of the variance in observed scores was universe score variance (i.e., variance attributable to persons). Variance attributable to occasions and slides was negligible, as was variance attributable to the persons-by-occasions and occasions-by-slides interactions. However, the persons-by-slides interaction

TABLE 4.1. Estimated Variance Components for Psychophysiological Protocol Data

Source of variation	*df*	Mean squares	Estimated variance component	% Total variance
Persons (*p*)	19	80.29	5.8941	47
Occasions (*o*)	1	40.92	0.2582	2
Slides (*s*)	5	17.37	0.1303	1
po	19	5.42	0.3183	3
ps	95	7.66	2.0697	17
os	5	8.02	0.2253	2
pos, e	95	3.52	3.5156	28

accounted for 17% of the variance, suggesting that individual respondents reacted differently to different slides. A relatively large residual variance component (28%) was also found, suggesting the presence of a significant amount of unmeasured or random error in the observed scores.

Table 4.2 presents the projection of the G-study data in Table 4.1 to four possible D-study designs, based on calculations analogous to the Spearman–Brown formula. The first column simply repeats the variance components from the G study. The last two rows in the first column provide the relative and absolute G coefficients (.50 and .47) for generalizing from a single slide on a single occasion. The next four columns show the impact of increasing number of slides from one to four or eight and increasing the number of occasions from one to two. For example, the second column presents the projected generalizability for a D study in which respondents watch four slides on one occasion. A substantial increase in generalizability is observed for this design, up to .77 for relative decisions and .74 for absolute decision.

Confirmatory Factor Analysis

Underlying the psychometric work on PTSD is the idea that PTSD is a hypothetical construct that is reflected in observed behaviors, including responses to assessment devices. One way of representing this notion is by means of factor analysis, a multivariate statistical procedure that can be used to identify the constructs or latent variables that appear to account for observed responses. The responses may be from a collection of items on a test or a collection of scores in a battery of tests. Viewing factor analysis from this perspective, one can think of the observed scores as being regressed on

TABLE 4.2. Alternative Decision Studies for Psychophysiological Protocol

	G study	Alternative D studies			
Source of variation	$N_o = 1$ $N_s = 1$	$N_o = 1$ $N_s = 4$	$N_o = 1$ $N_s = 8$	$N_o = 2$ $N_s = 4$	$N_o = 2$ $N_s = 8$
Persons (*p*)	5.8941	5.8941	5.8941	5.8941	5.8941
Occasions (*o*)	0.2582	0.2582	0.2582	0.1291	0.1291
Slides (*s*)	0.1303	0.0326	0.0163	0.0326	0.0163
po	0.3183	0.3183	0.3183	0.1591	0.1591
ps	2.0697	0.5174	0.2587	0.5174	0.2587
os	0.2253	0.0563	0.0282	0.0282	0.0141
pos, e	3.5156	0.8789	0.4394	0.4394	0.2197
Error variances					
Relative	5.9035	1.7146	1.0164	1.1160	0.6376
Absolute	6.5174	2.0617	1.3191	1.3059	0.7971
Generalizability coefficients					
Relative	.50	.77	.85	.84	.90
Absolute	.47	.74	.82	.82	.88

factors in a series of multiple regression equations, one for each of the items (or tests). Each equation takes this general form:

$$Y_{ji} = b_{Yi}F_{j1} + b_{Y2}F_{j2} + \ldots + b_{Ym}F_{jm} + e_{ji}$$

Here, let us consider each dependent variable (Y_{ji}) to be a score for person j on item i; the independent variables ($F_{j1}, F_{j2}, \ldots, F_{jm}$) are factor scores for person j; the regression weights ($b_{Y1}, b_{Y2}, \ldots, b_{Ym}$) are what are called the "factor loadings"; and e_{ji} symbolizes a residual uniqueness or error in prediction.

If, for example, there are 35 items (as on the Mississippi Scale), there will be 35 such equations, and each item score for an individual is a weighted sum of that person's standing on the factors, plus residual. As in any multiple regression procedure, the direction of influence is from the independent variables to the dependent variable. Thus, factor analysis highlights the idea that the factors or hypothetical constructs are responsible for the observed scores on the items. A large value for b would indicate that the latent variable has an important influence on an observed item score; conversely, a small value for this regression weight or factor loading would indicate that the latent variable does not seem to account for the observed score. In computing a factor-analytic solution for an observed data set, one is interested in the matrix of values for the factor loadings (the weights for the regressions of all items on the factors), a matrix depicting how the factors relate to one another, and possibly a matrix of residuals. Details on the computations of elements in these matrices may be found in many sources, including Harman (1976), Gorsuch (1983), and Loehlin (1992).

Historically, the more common factor-analytic strategy has been the unrestricted or exploratory approach. Using this approach, the factor analyst allows the data to determine the solution; that is, the data determine the values of the loadings of each item on each factor, frequently the number of factors, and under certain circumstances the extent to which factors intercorrelate with one another. The goal is to empirically discover the hypothetical constructs that are responsible for the observed pattern of relationships among the items. Sizes of loadings on the various resulting factors are examined; the content of items that load highest on each factor is evaluated; and the usual criterion for the adequacy of the exploratory solution is its interpretability vis-à-vis some a priori or post hoc conceptualization of the constructs.

Recently, emphasis has been given to what is called restricted or confirmatory factor analysis, in which features of the factor solution are hypothesized or specified beforehand. Based on a theoretical or conceptual framework, the factor analyst specifies the number of factors proposed to be responsible for the relationships within the data set. Furthermore, items are designated to load on particular factors; typically, each item is specified to load on only one factor to facilitate interpretation of the solution (An-

derson & Gerbing, 1988). Factors may or may not be allowed to correlate, depending upon the theory guiding the understanding of the constructs. The factor analyst also has control over how the residuals are to be treated; in most circumstances, they should not be permitted to covary. The adequacy of the hypothesized factor solution is evaluated by comparing a matrix of relationships among observed item scores, usually a variance–covariance matrix, with one that is derived using the hypothesized factor solution. Descriptions of the algorithms for estimating parameters for the hypothesized solution and for computing the reproduced matrix can be found in works by Joreskog and Sorbom (1979), Bollen (1989), and Loehlin (1992).

An important aspect of any hypothesized factor solution is that there be fewer parameter estimates (factor loadings, factor variances and covariances, and residuals) in the hypothesized solution than there are variances and covariances among the observed variables. When items are specified to load on only one factor, they are obviously specified not to load on the other factors, suggesting a series of potentially disconfirmable propositions about the underlying structure of the data. As Mulaik and James (1995) pointed out, the larger the difference between the number of parameter estimates and the number of variances and covariances among the observed variables (i.e., the more degrees of freedom), the greater the potential for disconfirming a model. "Good models" are those with a strong potential to be disconfirmed (many degrees of freedom) but for which the accompanying data do not support disconfirmation.

In the end, a discrepancy index serves as a basis of evaluation of a hypothesized solution. This index is a weighted sum of squared deviations of the reproduced associations from the observed associations. Ideally, the discrepancy value should be a minimum, as this outcome would indicate "close fit" or that the hypothesized constructs are likely responsible for the observed scores. Large discrepancy values suggest the possibility of an alternative representation of the factor structure. When the observed scores are continuous and multivariate normally distributed, and when either maximum likelihood or generalized least-squares estimation is used, the distribution of the discrepancy statistic takes on a known form, that of noncentral chi-square, with degrees of freedom equal to the difference between the number of parameter estimates in the hypothesized factor-analytic model and the number of variances and covariances among the observed variables. Low values of chi-square, relative to the degrees of freedom, suggest endorsement of the hypothesized factor structure. Discussion of the nature of the noncentral chi-square and its implications for evaluation of model–data fit is provided by Browne and Cudeck (1993).

Due to certain problems with the chi-square statistic as an indicator of model–data fit (Bentler & Bonett, 1980), many other fit indices have been developed in recent years. For overviews of the various indices, see Chapter 4 of Joreskog and Sorbom's (1993) SIMPLIS guide and the review by Hu and Bentler (1995).

A psychometric study by King and King (1994) illustrates the application of confirmatory factor analysis to PTSD assessment. The data were responses to the 35-item Mississippi Scale from 2,272 Vietnam theater and era veterans who participated in the NVVRS (Kulka et al., 1990). Consistent with Bagozzi and Heatherton's (1994) total disaggregation conceptualization of a multifaceted construct, PTSD was proposed as a second-order umbrella factor that subsumed four symptom categories or first-order factors, namely Reexperiencing and Situational Avoidance, Withdrawal and Numbing, Arousal and Lack of Control, and Guilt and Suicidality. Each of these first-order factors, in turn, was postulated to account for the responses to Mississippi Scale items, with content judged to reflect that symptom category. Residuals were specified to be uncorrelated. The hypothesized structure is depicted in Figure 4.4, where arrows indicate direction of influence. Because the data were considered ordinal and judged not multivariate normally distributed, a matrix of polychoric correlations was analyzed using weighted least-squares estimation (Browne, 1984).

To appraise the viability of the hypothesized structure, King and King (1994) employed sequential chi-square difference testing of hierarchically nested factor models. This process involves the systematic inspection of outcomes when several reasonable and increasingly restrictive models are fit to the data and compared with one another. One of the models is the hypothesized structure. Initially, they fit a four-factor first-order solution, each factor coinciding with a symptom category and allowing all four factors to covary (see Figure 4.5). This model is a multidimensional representation of PTSD in that it lacks the global organizing PTSD second-order factor that is ap-

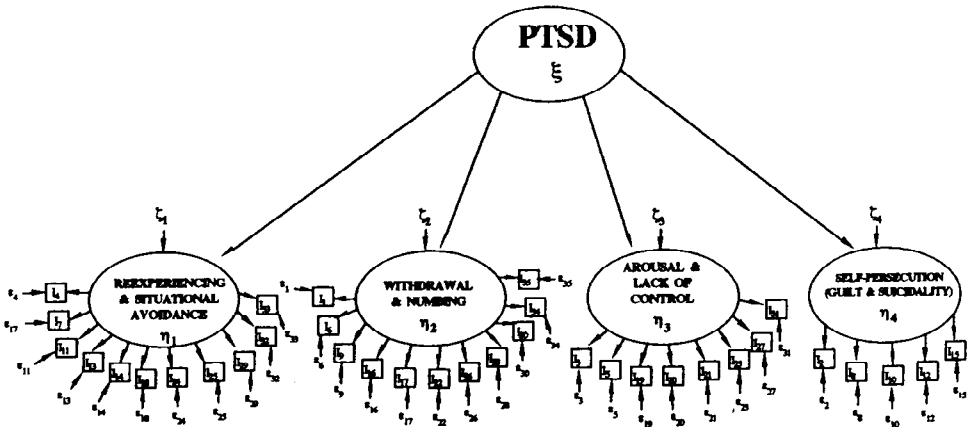
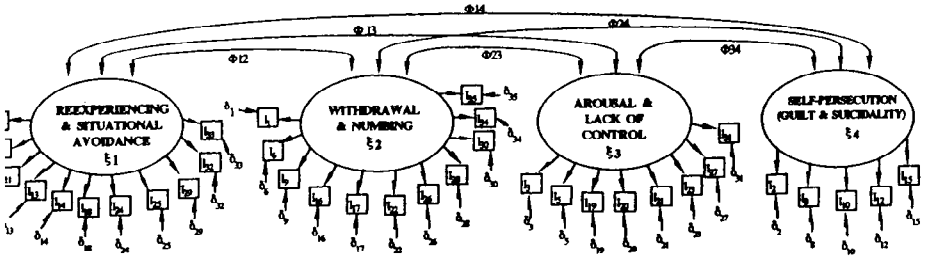


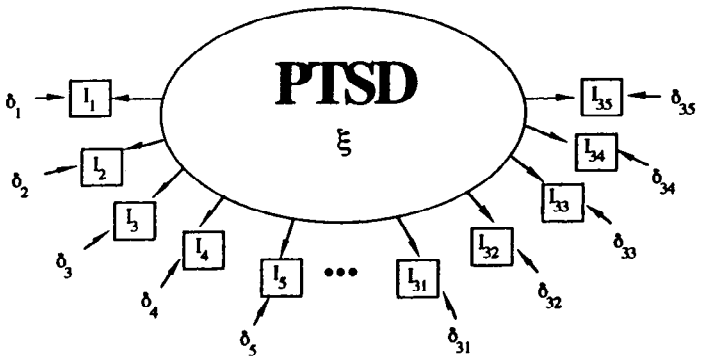
FIGURE 4.4. Hypothesized second-order model using LISREL (Joreskog & Sorbom, 1993) notation. I_j = Mississippi Scale item number. Reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from *Assessment*, Vol. 1, No. 3. Copyright 1994 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.



GURE 4.5. Four-factor first-order model, using LISREL (Joreskog & Sorbom, 1993) tation. I_j = Mississippi Scale item number. Reproduced by special permission of e Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from *Assment*, Vol. 1, No. 3. Copyright 1994 by PAR, Inc. Further reproduction is prohibitwithout permission from PAR, Inc.

ırent in the hypothesized model. Next, they fit the hypothesized second-der model (Figure 4.4). Finally, they fit a single-factor first-order solution, th all items loading on a unitary PTSD factor (see Figure 4.6).

Table 4.3 summarizes the confirmatory factor-analytic findings. Of par-ıular interest are the values associated with the chi-square difference tests ıbeled Dc^2). In a series of hierarchically nested models, the difference in ·o chi-square statistics is distributed as chi-square with the degrees of free-)m equal to the difference in the degrees of freedom associated with each odel (Steiger, Shapiro, & Browne, 1985). Therefore, the tests of the sigficance of the difference between chi-squares (again, the columns under :2) provide a mechanism for judging the competing models and assessing



IGURE 4.6. Single-factor first-order model, using LISREL (Joreskog & Sorbom,)93) notation. I_j = Mississippi Scale item number. Reproduced by special permison of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, om *Assessment*, Vol. 1, No. 3. Copyright 1994 by PAR, Inc. Further reproduction prohibited without permission from PAR, Inc.

TABLE 4.3. Chi-Square Difference Tests of Competing Models of Mississippi Scale Structure

Model	c^2	*df*	Dc^2	*df*	*p*
Four-factor first-order	1,023.27	554			
Single-factor second-order	1,024.35	556	1.08	2	.58
Single-factor first-order	1,051.41	560	28.14	4	< .001

Note. Findings are from King and King (1994).

whether the hypothesized model is most appropriate, given the data. As shown in Table 4.3, the first two models, the four-factor first-order solution and the single-factor second-order solution, were comparable in fit ($p = .58$). But the single-factor first-order solution differed from the other two (for both, $p < .001$). Hence, the hypothesized single-factor second-order solution is superior to the four-factor first-order solution because it achieves like fit with two fewer parameters being estimated. Moreover, the hypothesized structure appears superior to the single-factor first-order solution because the reduction in the number of parameter estimates in the latter model (four) produces unacceptable damage to fit. This is indicated by the large chi-square difference relative to the associated degrees of freedom.

In summary, confirmatory factor analysis enhances theory testing and development by allowing the researcher to statistically evaluate the soundness of a hypothesized factor structure. In addition, the parameter estimates that result are efficient, meaning that their standard errors are as small as they can be and therefore are presumably better approximations of the true parameter values. Also, the available statistical software packages furnish a wealth of detailed information regarding how models might be improved. Morris, Bergan, and Fulginiti (1991) offer interesting commentary on the use of confirmatory factor analysis versus exploratory factor analysis in clinical assessment research.

Item Response Theory

Item response theory (IRT) is a contemporary approach to test development and evaluation that has tended to supplant classical test theory in recent years. Like factor analysis, IRT presupposes that an individual's standing on a hypothetical construct or latent variable can predict how that person responds to an item on a psychometric instrument. Whereas the mathematical foundation for factor analysis is linear regression, the form of the relationship in IRT is that of a curvilinear probability function, portraying what is called an "item characteristic curve." In its early stage of development, the mathematical expression was the cumulative normal ogive, an S-shaped curve framed by an *X* axis representing scores or standing on the hypothetical construct or attribute of interest, and a *Y* axis representing the probability of

responding to a binary or dichotomous ("correct"/"incorrect") item in the keyed direction. As the theory was developed further, the normal ogive was replaced by a more mathematically tractable, logistic function (Birnbaum, 1968). The specific form of this function is governed by up to three parameters or item characteristics: (1) the item discrimination index, a reflection of how well the item distinguishes between those low and those high on the attribute, where low and high are determined by (2) an item difficulty or threshold level, the point on the attribute continuum at which the probability of choosing one response option over another is .50, and (3) a guessing parameter, the probability of guessing correctly or responding in the keyed direction when the person actually possesses little or none of the attribute. One advantage of IRT over classical test theory is that the estimates of these characteristics are not dependent upon the makeup of the sample upon which they are computed.

A researcher using IRT must first select the particular model that is most suited to the measure, with concern for item type, the dimensionality of the construct, and which blend or mixture of the possible item characteristics are needed to best capture the nature of the construct. The basic IRT model for simple, dichotomous items has been expanded to accommodate other forms of items. Masters and Wright (1984) and Thissen and Steinberg (1986) present details on the various available models, their relationships to one another, and applications, and excellent introductory presentations of the foundations of IRT are provided by Hulin, Drasgow, and Parsons (1983) and Hambleton, Swaminathan, and Rogers (1991).

Of particular interest here are items with multiple, graded response options, such as the Likert-type scales used for the Mississippi Scale, CAPS, PSS, LASC, and other PTSD assessment instruments. One IRT approach for such item types is Samejima's (1969) graded response model, which was employed by King, King, Fairbank, Schlenger, and Surface (1993) to study the items comprising the Mississippi Scale, with data again drawn from the NVVRS. Using one discrimination parameter and $k-1$ difficulty parameters (where k equals the number of options on the response scale), the graded response model provides for k operating characteristics. These are like item characteristic curves for each response option, each depicting the probability that an individual will select that option as a function of his or her standing on the attribute. An example from the King et al. findings, operating characteristics for Mississippi Scale Item 6 ("I am able to get emotionally close to others"), is presented as Figure 4.7. As can be seen, for this item, optimal probabilities of the five possible responses are well distributed across the range of the attribute continuum. In addition, only those individuals who are quite high on the PTSD dimension have a reasonable probability of endorsing the "never" option.

More important, IRT applications supply a plot of an item's reliability in the form of an item information curve, depicting the amount of information or precision of measurement (Y axis) across the range of possible attri-

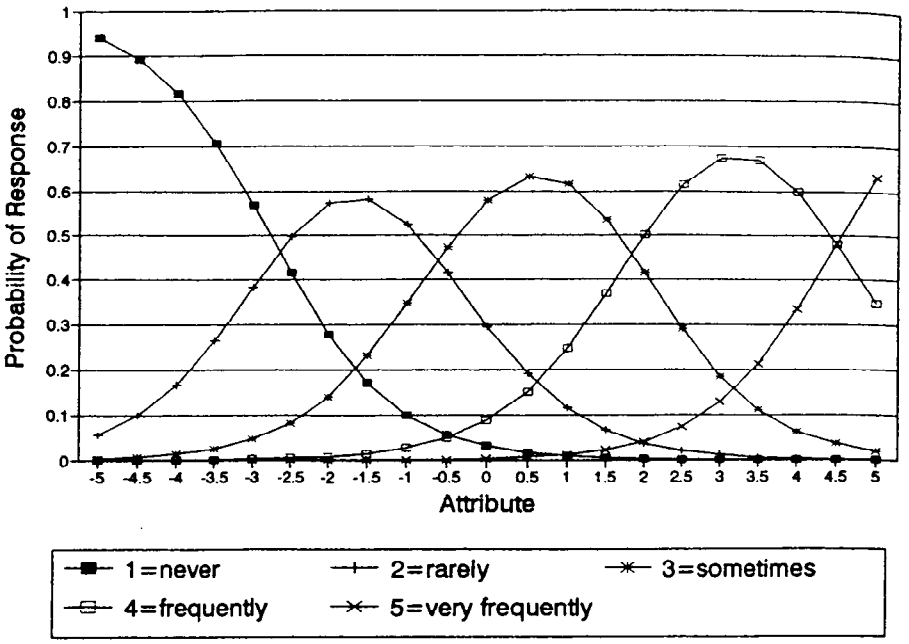


FIGURE 4.7. Operating characteristics for Mississippi Scale Item 6. Adapted from King, King, Fairbank, Schlenger, and Surface (1993, p. 462). Copyright 1994 by the American Psychological Association. Adapted by permission.

bute scores (*X* axis). The item information curve can be interpreted as item reliability since, for a given position on the attribute continuum, the square root of its inverse is the standard error of measurement. For example, as King et al. (1993) noted, and as demonstrated in Figure 4.8, Mississippi Scale Item 8 ("... I wish I were dead") is highly accurate or precise in discriminating among persons very high on the PTSD dimension, compared to Item 24 ("I fall asleep easily at night"), which provides moderate information or precision across a broad attribute range. A similar test information function may also be derived to portray how the full collection of items performs across the range of attribute positions.

In addition to supplying estimates of item parameters, item characteristic curves, item information functions, and the test information function—all of which assist in understanding the items as indicators of the construct—IRT also generates maximum likelihood estimates of individuals' scores. An important point is that these scores are item invariant. In other words, it is possible to administer different and varying numbers of items drawn from a common content bank to two or more individuals and derive scores that share the same attribute scale. In fact, one of the more valuable applications of IRT has been the implementation of individual computerized adaptive testing, in which an examinee is administered only as many items as are neces-

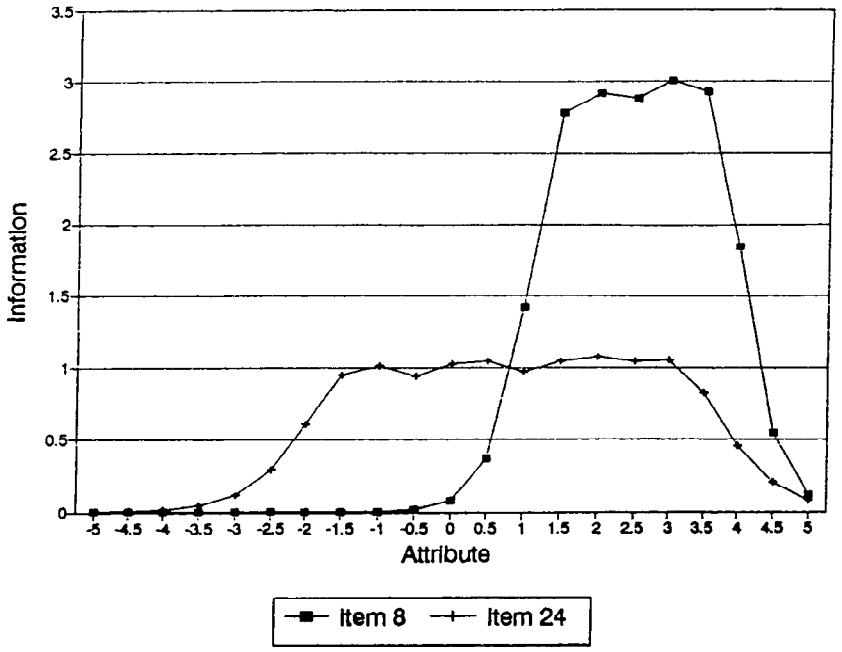


FIGURE 4.8. Item information functions for Mississippi Scale Items 8 and 24. Adapted from King, King, Fairbank, Schlenger, and Surface (1993, pp. 465–466). Copyright 1994 by the American Psychological Association. Adapted by permission.

sary to determine his or her standing on an attribute at some predetermined level of precision. Largely used, to date, in cognitive assessment (primarily aptitude tests for admission to college and professional programs), computer-administered, individually tailored tests appear to have great potential for affective measurement as well, such as the assessment of PTSD and other diagnostic entities. Wainer (1991) offers an excellent edited volume on computerized adaptive testing.

We previously stated that IRT-based item characteristics are sample invariant; that is, the discrimination, difficulty, and guessing parameters should be the same, regardless of the sample used to estimate them. Therefore, if one discovers that parameter estimates are not the same from population to population, then the possibility of differential item functioning or item bias arises. In PTSD research, for example, differential item functioning across groups might suggest that the item in question has different meanings for the groups, and hence that PTSD may not be manifest in a consistent manner across trauma populations. Another application of differential item functioning techniques is for equating test items across language translations. To the extent that equivalent item characteristics can be documented, one has more confidence in comparable PTSD measurement across

language groups, thereby enhancing the potential validity of cross-cultural research. King et al. (1993) give further specifics on the implications of IRT for PTSD assessment.

SUMMARY AND CONCLUSIONS

In the last 15 years, extraordinary progress has been made in the development of reliable and valid PTSD assessment tools. The field of PTSD assessment has rapidly evolved from the situation in the early 1980s, when no measures of any kind existed, to the present situation, in which more than two-dozen questionnaires, interviews, and physiological protocols are available for a wide variety of assessment tasks. These instruments meet the practical needs of clinicians who require psychometrically sound tools for diagnosing and assessing individuals with PTSD and monitoring their progress in treatment. Perhaps even more important, they have also greatly facilitated empirical research regarding the clinical phenomenology and basic psychopathological processes involved in the response to traumatic life events.

In developing and evaluating new PTSD assessment instruments, investigators have generally followed the psychometric principles and techniques outlined in this chapter. We identified five basic stages, each involving multiple tasks. These stages include specifying the purpose of the instrument, defining the construct, designing the instrument, pilot testing and revising, and establishing reliability and validity. Throughout the entire process there is considerable interplay between creative input, based on a clear conceptualization of PTSD, and empirical feedback, based on rigorous investigation of the statistical properties of the instrument. To date, much of the statistical evaluation of PTSD instruments has been based on the methods of classical test theory. The contemporary approaches we described, including generalizability theory, confirmatory factor analysis, and item response theory, should greatly enhance the revision of existing instruments and the development of new, more sophisticated PTSD measures.

Although much progress has been made, there are several issues that still need to be addressed. First, with few exceptions, most PTSD assessment instruments are quite face valid and do not include items to assess response bias. This leaves them subject to deliberate distortion of responses, which can markedly reduce their usefulness. PTSD is not unique in this respect; most self-report measures of other types of psychopathology suffer from the same limitation. Nonetheless, the detection and quantification of response bias is an important area that deserves consideration by PTSD test developers. Second, many PTSD instruments were developed in a particular trauma population, such as combat veterans or rape victims, and have not been evaluated across populations. This creates a confounding of instrument and population that limits direct comparisons of PTSD symptomatology resulting from

different types of trauma. A related concern is that many PTSD instruments were initially developed on relatively small samples of convenience, and few have been normed on large, representative samples. It will be important in the future to adopt a standard battery of PTSD instruments that can be used across different traumatized populations, and to develop adequate norms for each measure.

Third, much more work is needed on the validation of existing PTSD instruments. For example, although there is ample evidence supporting the convergent validity of various PTSD measures, little has been documented regarding discriminant validity; that is, test developers often report robust correlations between different measures of PTSD, but they typically fail to report the equally necessary weak correlations between measures of PTSD and measures of other constructs. Such evidence is crucial for demonstrating that scores on PTSD scales reflect "PTSD-ness" and are not better conceptualized as measures of anxiety or depression, for example, or as measures of global distress. Finally, as the PTSD construct evolves and as the database on existing instruments accumulates, it will become necessary to revise specific items or even entire measures to address empirically identified weaknesses and thereby enhance reliability and validity. Such revisions, such as the recent revision of the MMPI, can be unsettling for clinicians and researcher alike, but they are essential for continued progress in the field.

REFERENCES

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling: A Multidisciplinary Journal, 1,* 35–67.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E., & Erbaugh, J. K. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (Part 5, pp. 397–474). Reading, MA: Addison-Wesley.

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Klauminzer, G., Charney, D. S., & Keane, T. M. (1990). A clinician rating scale for assessing current and lifetime PTSD: The CAPS-1. *Behavior Therapist, 13,* 187–188.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City: American College Testing Program.

Briere, J. (1995). *Trauma Symptom Inventory (TSI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Browne, M. W. (1984). Asymptotic distribution free methods in analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carroll, K. M. (1995). Methodological issues and problems in the assessment of substance abuse. *Psychological Assessment, 7,* 349–358.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York: HarperCollins.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Davidson, J. R. T., & Foa, E. B. (1991). Diagnostic issues in posttraumatic stress disorder: Considerations for the DSM-IV. *Journal of Abnormal Psychology, 100,* 346–355.

Davidson, J. R. T., Smith, R. D., & Kudler, H. S. (1989). Validity and reliability of the DSM-III criteria for post-traumatic stress disorder: Experience with a structured interview. *Journal of Nervous and Mental Disease, 177,* 336–341.

Derogatis, L. R. (1983). *SCL-90-R: Administration, scoring, and procedures manual—II.* Towson, MD: Clinical Psychometric Research.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Foa, E. B., Riggs, D. S., Dancu, C. V., & Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of Traumatic Stress, 6,* 459–473.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Philadelphia: Saunders.

Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. K. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hammarberg, M. (1992). Penn Inventory for Posttraumatic Stress Disorder: Psychometric properties. *Psychological Assessment, 4,* 67–76.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed., rev.). Chicago: University of Chicago Press.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Herman, J. L. (1992). *Trauma and recovery.* New York: Basic Books.

Holden, R. R., & Fekken, G. C. (1990). Structured psychopathological test item characteristics and validity. *Psychological Assessment, 2,* 35–40.

Horowitz, M. J., Wilner, N., & Alvarez, W. (1979). Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine, 41,* 209–218.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (Part 5, pp. 76–100). Thousand Oaks, CA: Sage.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Jackson, D. N. (1976). *The Basic Personality Inventory.* Port Huron, MI: Research Psychologists Press.

Janoff-Bulman, R. (1992). *Shattered assumptions: Towards a new psychology of trauma.* New York: Free Press.

Joreskog, K. G., & Sorbom, D. (1979). *Advances in factor analysis and structural equation models.* Lanham, MD: University Press of America.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, NJ: Erlbaum.

Kashy, D. A., & Snyder, D. K. (1995). Measurement and data analytic issues in couples research. *Psychological Assessment, 7,* 338–348.

Keane, T. M., Caddell, J. M., & Taylor, K. L. (1988). Mississippi Scale for Combat-Related Post-Traumatic Stress Disorder: Three studies in reliability and validity. *Journal of Consulting and Clinical Psychology, 56,* 85–90.

Keane, T. M., Malloy, P. F., & Fairbank, J. A. (1984). Empirical development of an MMPI subscale for the assessment of combat-related posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology, 52,* 888–891.

Keane, T. M., Wolfe, J., & Taylor, K. L. (1987). Post-traumatic stress disorder: Evidence for diagnostic validity and methods of psychological assessment. *Journal of Clinical Psychology, 43,* 32–43.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399.

King, D. W., King, L. A., Fairbank, J. A., Schlenger, W. E., & Surface, C. R. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Post-Traumatic Stress Disorder: An application of item response theory. *Psychological Assessment, 5,* 457–471.

King, D. W., King, L. A., Leskin, G., & Foy, D. W. (1995). The Los Angeles Symptom Checklist: A self-report measure of post-traumatic stress disorder. *Assessment, 2,* 1–17.

King, L. A., & King, D. W. (1994). Latent structure of the Mississippi Scale for Combat-Related Post-traumatic Stress Disorder: Exploratory and higher-order confirmatory factor analyses. *Assessment, 1,* 275–291.

Koretzky, M. B., & Peck, A. H. (1990). Validation and cross-validation of the PTSD Subscale of the MMPI with civilian trauma victims. *Journal of Clinical Psychology, 46,* 296–300.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines.* Newbury Park, CA: Sage.

Kulka, R. A., Schlenger, W. E., Fairbank, J. A., Hough, R. L., Jordan, B. K., Marmar,

C. R., & Weiss, D. S. (1990). *Trauma and the Vietnam War generation: Report on the findings from the National Vietnam Veterans Readjustment Study.* New York: Brunner/Mazel.

La Rue, A., & Markee, T. (1995). Clinical assessment research with older adults. *Psychological Assessment, 7,* 376–386.

Litz, B. T. (1992). Emotional numbing in combat-related post-traumatic stress disorder: A critical review and reformulation. *Clinical Psychology Review, 12,* 417–432.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lyons, J. A., Caddell, J. M., Pittman, R. L., Rawls, R., & Perrin, S. (1994). The potential for faking on the Mississippi Scale for Combat-Related PTSD. *Journal of Traumatic Stress, 7,* 441–445.

Lyons, J. A., & Keane, T. M. (1992). Keane PTSD Scale: MMPI and MMPI-2 update. *Journal of Traumatic Stress, 5,* 111–117.

Malloy, P. F., Fairbank, J. A., & Keane T. M. (1983). Validation of a multimethod assessment of posttraumatic stress disorders in Vietnam veterans. *Journal of Consulting and Clinical Psychology, 51,* 488–494.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

McCann, I. L., & Pearlman, L. A. (1990). *Psychological trauma and the adult survivor: Theory, therapy, and transformation.* New York: Brunner/Mazel.

Morris, R. J., Bergan, J. R., & Fulginiti, J. V. (1991). Structural equation modeling in clinical assessment research with children. *Journal of Consulting and Clinical Psychology, 59,* 371–379.

Mulaik, S. A., & James, L. R. (1995). Objectivity and reasoning in science and structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (Part 7, pp. 118–138). Thousand Oaks, CA: Sage.

Newman, E., Kaloupek, D. G., & Keane, T. M. (1996). Assessment of posttraumatic stress disorder in clinical and research settings. In B. A. van der Kolk, A. C. McFarlane, & L. Weisaeth (Eds.), *Traumatic stress: The effects of overwhelming experiences on mind, body, and society* (pp. 242–275). New York: Guilford Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Okazaki, S., & Sue, S. (1995). Methodological issues in assessment research with ethnic minorities. *Psychological Assessment, 7,* 367–375.

Orsillo, S., Weathers, F. W., Litz, B. T., Steinberg, H. R., Huska, J. A., & Keane, T. M. (1996). Current and lifetime psychiatric disorders among veterans with war-zone-related post-traumatic stress disorder. *Journal of Nervous and Mental Disease, 184,* 307–313.

Pitman, R. K., Orr, S. P., Forgue, D. F., de Jong, J. B., & Claiborn, J. M. (1987). Psychophysiologic assessment of posttraumatic stress disorder imagery in Vietnam combat veterans. *Archives of General Psychiatry, 44,* 970–975.

Prigatano, G. P., Parsons, O. A., & Bortz, J. J. (1995). Methodological considerations in clinical neuropsychological research: 17 years later. *Psychological Assessment, 7,* 396–403.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph No. 17, 34*(4, Pt. 2).

Saunders, B. E., Arata, C. M., & Kilpatrick, D. G. (1990). Development of a crime-related posttraumatic stress disorder scale for women within the Symptom Checklist-90—Revised. *Journal of Traumatic Stress, 3,* 439–448.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24,* 399–411.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1990). *Structured Clinical Interview for DSM-III-R.* Washington, DC: American Psychiatric Association Press.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50,* 253–263.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Erlbaum.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.

Vreven, D. L., Gudanowski, D. M., King, L. A., & King, D. W. (1995). The civilian version of the Mississippi PTSD Scale: A psychometric evaluation. *Journal of Traumatic Stress, 8,* 91–109.

Wainer, H. (1991). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

Watson, C. G., Juba, M. P., Manifold, V., Kucala, T., & Anderson, P. E. D. (1991). The PTSD Interview: Rationale, description, reliability, and concurrent validity of a DSM-III-based technique. *Journal of Clinical Psychology, 47,* 179–188.

Weathers, F. W., Blake, D. D., Krinsley, K. E., Haddad, W. H., Huska, J. A., & Keane, T. M. (1992, November). *The Clinician-Administered PTSD Scale: Reliability and construct validity.* Paper presented at the annual meeting of the Association for Advancement of Behavior Therapy, Boston, MA.

Weathers, F. W., Haddad, W. P., Litz, B. T., Keane, T. M., Palmieri, P. A., & Steinberg, H. S. (1995, November). *Intimacy in war-zone-related PTSD: Cognitive and affective responses to simulated interpersonal situations.* Paper presented at the annual meeting of the Association for Advancement of Behavior Therapy, Washington, DC.

Weathers, F. W., & Litz, B. T. (1994). Psychometric properties of the Clinician-Administered PTSD Scale-Form 1 (CAPS-1). *PTSD Research Quarterly, 5,* 2–6.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., & Keane, T. M. (1993, October). *The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility.* Paper presented at the annual meeting of the International Society for Traumatic Stress Studies, San Antonio, TX.

Weathers, F. W., Litz, B. T., Herman, D. S., Keane, T. M., Steinberg, H. R., Huska, J. A., & Kraemer, H. C. (1996). The utility of the SCL-90-R for the diagnosis of war-zone-related PTSD. *Journal of Traumatic Stress, 9,* 111–128.

Weiner, I. B. (1995). Methodological considerations in Rorschach research. *Psychological Assessment, 7,* 330–337.